Kombinasi Ekstraksi Kata Kunci dan Ekspansi Kueri Untuk Deteksi Isu Etik pada Ringkasan Penelitian Kesehatan

Combination of Keyword Extraction and Query Expansion to Detect Ethical Issues in Health Research Summary

Mohammad Zaenuddin Hamidi[1], Diana Purwitasari[2], Ratih Nur Esti Anggraini[3]
[1,2,3]Magister Teknik Informatika, Institut Teknologi Sepuluh Nopember
E-mail: [1]mzaenuddinhamidi@gmail.ac.id, [2]diana@if.its.ac.id, [3]ratih_nea@if.its.ac.id
*penulis korespondensi: diana@if.its.ac.id

Abstrak

Penelitian kesehatan harus melalui proses telaah etik yang bertujuan untuk mengantisipasi dugaan atas risiko fisik, sosial, ekonomi dan psikologis. Secara etik penelitian kesehatan dapat diterima apabila pada penelitian tersebut mampu dibuktikan dengan metode ilmiah yang valid serta lulus uji etik sebelum penelitian dilakukan. Untuk memastikan pada ringkasan penelitian kesehatan terdapat aspek etik, dibutuhkan kata kunci yang dapat dijadikan representasi dari isi ringkasan tersebut. Salah satu pendekatan yang sering dilakukan adalah dengan menghitung frekuensi kemunculan kata dalam dokumen. Pendekatan lain yaitu pendekatan YAKE dan keyBERT yang tidak hanya menghitung frekuensi kata namun juga menghitung konteks kata. Selain melakukan ekstraksi dilakukan juga proses ekspansi kueri sebagai upaya memperluas istilah yang dapat mewakili masing-masing aspek etik. Salah satu pendekatan yang digunakan untuk ekspansi kueri adalah model word2vec. Penelitian ini mengusulkan metode deteksi aspek etik pada ringkasan penelitian kesehatan dengan ekspansi kueri dengan word2vec dan proses ekstraksi kata kunci dengan TFIDF, YAKE dan keyBERT yang dikombinasikan dengan fuzzy. Hasil eksperimen menunjukkan bahwa metode paling unggul secara presisi yaitu YAKE dan gabungan antara TFIDF + YAKE + keyBERT dengan nilai tertinggi 46% kemudian untuk recall model YAKE mendapat nilai tertinggi dengan angka 72% dan untuk nilai F1-Score yang paling unggul adalah metode YAKE dengan nilai tertinggi 54%.

Kata kunci: Ekspansi Kueri, Ekstraksi Kata Kunci, Fuzzy, Isu Etik

Abstract

Health research must go through an ethical review process to anticipate allegations of physical, social, economic, and psychological risks. Health research can be ethically accepted if it can be proved by a valid scientific method and passes an ethical test before it is carried out. To ensure that, health research summaries must contain ethical aspects, and keyword has needed that can be used as representations the contents of the resume. One approach that is frequently used is to calculate the frequency of occurrence of words in documents. Another approach is the YAKE and keyBERT approaches, which not only that but also word context. In addition to extracting, a query expansion process is also carried out to broaden the terms that can represent each ethical aspect. One approach used for query expansion is the word2vec model. This study proposes a method of detecting ethical aspects in health research summaries by query expansion with word2vec and keyword extraction processes with TFIDF, YAKE and keyBERT combined with fuzzy. The experimental results show that the most superior methods in term of precision is YAKE and the combination of TFIDF + YAKE + keyBERT with the highest score of 46%. For recall, the YAKE model gets the highest score with 72%, and for the F1-Score, the most superior method is the YAKE method with the highest value of 54%.

Keywords: Query Expansion, Keyword Extraction, Fuzzy, Ethics Issues

1. PENDAHULUAN

Perkembangan ilmu pengetahuan ditandai dengan meningkatnya penelitian pada berbagai bidang tidak terkecuali penelitian pada bidang kesehatan. Terdapat 29 juta artikel jumlah publikasi dalam bidang medis sejak awal Januari tahun 2019 [1]. Perkembangan teks pada bidang kesehatan menjadi sebuah tantangan topik penelitian pemrosesan teks (*text processing*). Suatu penelitian dapat diterima secara etis jika dapat dibuktikan dengan metode ilmiah yang valid. Sebuah studi yang tidak sehat secara ilmiah berisiko kerugian atau tidak mendapatkan manfaat nilai ekstrinsik dan nilai intrinsik bagi ilmu pengetahuan. Telaah etik diperlukan untuk memastikan proses penelitian yang dilakukan sesuai dengan aturan etik serta tidak ada pelanggaran prinsip dasar terutama terhadap hak subjek penelitian. Kriteria etik dibagi menjadi tujuh standar etik dalam penelitian kesehatan yaitu: nilai sosial, ilmiah, pemerataan beban manfaat, potensi risiko, bujukan (*inducements*), perlindungan privasi dan persetujuan setelah penjelasan [2]. Sifat jujur, sopan, memperhatikan kerahasiaan dan bertanggung jawab pada subyek penelitian merupakan bagian dari aspek etik penelitian kesehatan[3].

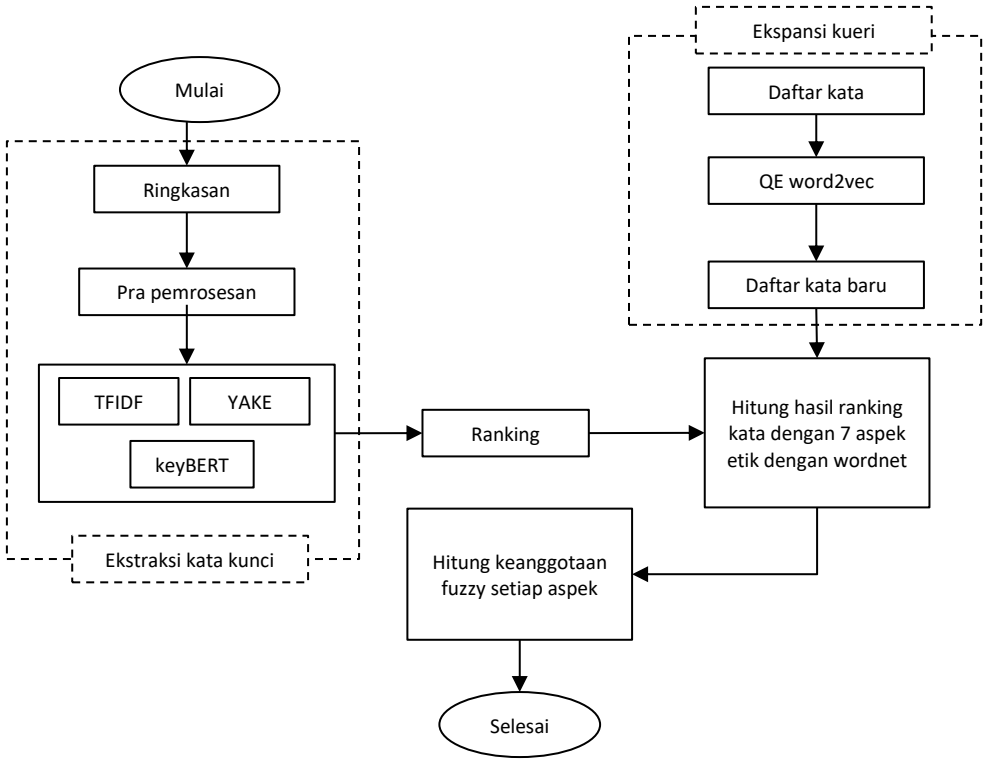
Perkembangan teknologi saat ini memungkinkan proses telaah etik dapat terkomputerisasi. Penelitian tentang bidang *text mining* dapat dilakukan sehingga kode etik dari sebuah ringkasan penelitian kesehatan terpenuhi berdasarkan kriteria yang ada [4]. *Text mining* dikenal sebagai proses penambangan data teks atau pencarian pengetahuan dari basis data tekstual yang mengacu pada proses penggalian pola atau pengetahuan dari dokumen teks. [5]. Penelitian terkait aspek *base* biasanya digunakan untuk melakukan klasifikasi seperti yang dilakukan oleh [6] dan kategorisasi teks yang memanfaatkan teks ulasan pengguna sebagai aspek kategorisasi [7]. namun pada penelitian ini aspeknya adalah tujuh standar etik penelitian kesehatan. Pada penelitian ini, kata kunci dari dokumen penelitian kesehatan digunakan untuk diekstraksi pada proses telaah etik. Proses mendeteksi aspek etik dalam sebuah ringkasan penelitian kesehatan membutuhkan kata kunci yang mewakili setiap aspek. Untuk meningkatkan fungsionalitas penerapan kata kunci banyak digunakan pada sistem temu kembali informasi [8]. Pemanfaatan kata kunci sebagai pencocokan dokumen dinilai mampu bekerja dengan baik dengan penilaian yang dilakukan oleh manusia [9].

Umumnya, metode yang digunakan untuk melakukan ekstraksi kata kunci adalah *term frequency inverse document frequency* (TF-IDF) [10]. TF-IDF memanfaatkan frekuensi kemunculan kata dalam teks untuk menghasilkan kata kunci artinya semakin banyak kata tersebut muncul maka potensi kata tersebut menjadi kata kunci semakin besar. Jika melihat bentuk data teks ringkasan, penggunaan TF-IDF untuk proses ekstraksi kata kunci dinilai tidak cukup. Hal tersebut dikarenakan proses perhitungannya hanya memanfaatkan frekuensi kemunculan kata. Untuk itu perlu dicoba pendekatan lain selain TF-IDF misalnya YAKE atau keyBERT. YAKE merupakan salah satu pendekatan ekstraksi kata kunci yang tidak hanya melakukan proses ekstraksi kata kunci berdasarkan frekuensi kemunculan kata tetapi faktor lain seperti posisi, awalan, konteks dan keberagaman kata juga dapat dipertimbangkan sebagai kandidat kata kunci [11][12]. Sementara itu, keyBERT merupakan suatu *library* yang memanfaatkan *transformer* dan *cosine similarity* untuk mengekstrak kata kunci dan frasa kunci secara efisien yang mewakili dokumen tertentu dengan lebih baik [13]. Jika dilihat dari konsepnya TFIDF memanfaatkan frekuensi kata, sementara keyBERT memanfaatkan konteks kata sedangkan YAKE memanfaatkan keduanya untuk itu pada penelitian ini dicoba menggabungkan konsep tersebut untuk melihat seberapa efektif penggunaan frekuensi dan konteks kata terhadap kata kunci yang dihasilkan. TF-IDF, YAKE dan beberapa metode lain yang memanfaatkan kesamaan topik untuk melakukan ekstraksi kata kunci [14]. Baik TF-IDF, keyBERT dan YAKE merupakan model yang menggunakan pendekatan *unsupervised*. Meskipun model *unsupervised* lebih baik dibandingkan dengan *suvervised* namun model *unsupervised* tidak menjamin dapat mewakili isi ringkasan hal ini disebabkan karena kurangnya penggunaan informasi semantik. Sementara itu, analisis semantik biasanya menimbulkan ketidakjelasan dan ketidakpastian sehingga pendekatan fuzzy dapat digunakan untuk menyelesaikan masalah tersebut [14]. Dengan mengadopsi fungsi keanggotaan penerapan fuzzy mampu memperbaiki masalah ketidakpastian [15].

Untuk melakukan pencocokan dokumen, dibutuhkan kata atau kalimat acuan untuk mengetahui tingkat keterkaitan antar kata atau kalimat. Dalam proses telaah etik tentu dibutuhkan kamus kata yang luas yang mampu mewakili masing-masing ketujuh aspek etik. Untuk memperkaya atau memperluas kata dibutuhkan suatu metode salah satunya adalah dengan melakukan ekspansi kueri. Pendekatan ekspansi kueri digunakan sebagai upaya memperluas atau memperkaya kata dari masing-masing aspek etik yang terkandung dalam penelitian kesehatan. Secara umum ekspansi kueri terdiri dari tiga jenis yaitu statistik, semantik dan hibrid [16]. Terdapat beberapa metode untuk melakukan proses ekspansi kueri salah satunya adalah menggunakan *word embedding* [17]. *Word embedding* menggunkan *word2vec* untuk mengukur kesamaan antara kueri dan dokumen dengan memanfaatkan kesamaan semantik untuk mencari kandidat kata [18].

Penelitian ini mengusulkan metode pengembangan proses deteksi etik dengan metode ekstraksi kata kunci yang dikombinasikan dengan metode ekspansi kueri serta fuzzy. Ekstraksi kata kunci digunakan untuk melakukan proses ekstraksi kata kunci yang dapat mewakili isi ringkasan data. Metode fuzzy dimanfaatkan untuk memetakan tingkat kemiripan dari kata hasil ekstraksi kata kunci dan ekspansi kueri sesuai dengan kriteria standar etik penelitian kesehatan.

2. METODE PENELITIAN



Gambar 1 Metode Penelitian

Pada penelitian ini, terdapat tiga langkah penting yang dilakukan yaitu ekstraksi kata kunci, ekspansi kueri dan proses fuzzy seperti yang terlihat pada Gambar 1. Ekstraksi kata kunci digunakan untuk mengekstraksi kata kunci atau kata yang dianggap penting dalam penelitian kesehatan. Sebelum dilakukan proses ekstraksi kata kunci terlebih dahulu dilakukan pra pemrosesan data untuk membersihkan data. Terdapat tiga metode yang digunakan untuk melakukan proses ekstraksi kata kunci yaitu TF-IDF, YAKE dan keyBERT. Ekspansi kueri digunakan untuk memperluas kata atau istilah yang dapat mewakili tujuh aspek etik penelitian kesehatan. Proses ekspansi kueri dilakukan dengan metode *word2vec*. Hasil dari proses ekstraksi

kata kunci ini akan dicari aspek kemiripannya dengan hasil ekspansi kueri menggunakan model *wordnet similarity* dan terakhir adalah melakukan perhitungan fuzzy terhadap hasil proses pencarian kemiripan.

2.1 Sumber data

Data yang digunakan pada penelitian ini adalah data ringkasan penelitian kesehatan yang bersumber dari open repositori dari beberapa perguruan tinggi negeri yang ada di Indonesia salah satunya adalah Universitas Sumatra Utara (https://repositori.usu.ac.id/). Data yang digunakan berupa data ringkasan penelitian sebab data tersebut yang bersifat terbuka. Jumlah data yang digunakan pada penelitian ini sebanyak 174 data ringkasan penelitian. Selanjutnya data yang telah dikumpulkan akan dilakukan pelabelan oleh ahli yang dalam hal ini dilakukan oleh Komisi Etik Penelitian Kesehatan (KEPK) Universitas Mataram. Proses pelabelan dilakukan mengacu pada tujuh aspek penelitian kesehatan.

Tabel 1 Penjelasan masing-masing aspek etik

No	Aspek Etik	Definisi	Contoh Kalimat
1	Sosial	Penelitian yang dilakukan memiliki *novelty* atau nilai kebaruan dan memiliki manfaat untuk evaluasi kebijakan dalam dunia kesehatan.	Hasil dari penelitian ini dapat digunakan sebagai dasar untuk memberikan rekomendasi terkait kebijakan daerah maupun nasional dalam hal promosi kesehatan di bidang KIA dan memperbaiki akses serta utilitas pelayanan bagi ibu.
2	Ilmiah	Penelitian mampu menjelaskan metode, prosedur, jenis sampel yang digunakan dalam proses penelitian serta variabel dan definisi operasional.	Penelitian ini adalah analitik dengan desain *cross-sectional*. Pengumpulan data dilakukan dengan cara *simple random* sampling. Populasi penelitian adalah anak usia sekolah dasar yang memenuhi kriteria penelitian dan bersedia mengikuti seluruh prosedur penelitian
3	Pemerataan beban dan manfaat	Dalam penelitian tercantum beban yang merata pada proses distribusinya atau tidak ada diskriminasi terhadap subjek atau sampel. Tidak hanya beban melainkan manfaat juga harus dibagi secara merata dan adil.	Peneliti bersikap adil atau *justice* kepada seluruh partisipan dan tidak membedakan perlakuan yang diberikan. selain itu peneliti juga tidak membeda-bedakan perlakuan pada subjek baik itu yang tua maupun yang muda. semua diperlakukan sama dan adil selama proses penelitian berlangsung.
4	Risiko	Dalam penelitian menggambarkan bahwa risiko yang ditimbulkan dalam penelitian seminimal mungkin. Selain itu kerugian yang mungkin timbul dalam penelitian juga dapat digambarkan	lokasi pengambilan di pembuluh darah lengan kiri atau kanan yang dikerjakan oleh seseorang yang ahli dibidangnya, sehingga risiko yang mungkin timbul saat pengambilan darah akan sangat kecil
5	Bujukan	Dalam penelitian dijelaskan insentif yang dapat diterima oleh subyek baik berupa uang, pengobatan gratis, hadiah dan layanan lainnya yang bersifat gratis.	sebagai kompensasi atas penelitian ini maka peneliti memberikan hadiah kecil kepada subjek berupa cinderamata sebagai bentuk rasa terima kasih telah berkenan mengikuti dan menjadi subjek pada penelitian
6	Privasi	Peneliti mampu menjaga kerahasiaan dari subyek. Selian itu juga dijelaskan bagaimana proses atau cara menjaga rahasia subyek dalam penelitian terutama yang sifatnya data pribadi dan tidak untuk umum.	Kerahasiaan identitas responden dijaga oleh peneliti dan hanya digunakan untuk kepentingan penelitian, dengan cara memberikan kode atau tanda pada lembar kuesioner dan kode tersebut hanya diketahui oleh peneliti itu sendiri.
7	Persetujuan setelah penjelasan	Peneliti dapat menggambarkan bagaimana proses memberikan penjelasan kepada subyek yang berkaitan dengan tujuan, manfaat, kemungkinan risiko yang akan timbul dalam proses penelitian serta dampak yang mungkin bisa timbul akibat penelitian yang dilakukan. Apabila sudah dijelaskan peneliti mampu menggambarkan persetujuan dari subyek setelah dilakukan penjelasan.	Responden menyatakan setuju apabila bersedia menandatangani *informed consent* atau lembar persetujuan tersebut. Responden yang bersedia dapat menandatangani lembar persetujuan yang telah disediakan oleh peneliti, dan yang tidak bersedia menjadi calon responden berhak menolak dan mengundurkan diri

Tabel 1 merupakan penjelasan masing-masing aspek dan contoh-contoh kalimat yang

berkaitan dengan masing-masing aspek. Sebagai contoh jika dalam ringkasan penelitian terdapat kalimat "Kerahasiaan identitas responden dijaga oleh peneliti dan hanya digunakan untuk kepentingan penelitian, dengan cara memberikan kode atau tanda pada lembar kuesioner dan kode tersebut hanya diketahui oleh peneliti itu sendiri" maka yang dapat dijadikan kata kunci adalah "kerahasiaan" dan "kode" sehingga kalimat tersebut sudah mampu mewakili aspek privasi dari subyek penelitian.

2.2 Ekspansi kueri

Ekspansi kueri digunakan untuk memperkaya kata yang akan menjadi rujukan yang mewakili masing-masing aspek etik. Metode yang digunakan untuk melakukan ekspansi adalah pendekatan *word embedding word2vec*. *Word2vec* pada dasarnya adalah teknik penyematan kata yang digunakan untuk mengubah kata-kata dalam kumpulan data menjadi vektor sehingga mesin dapat memahaminya. Vektor kata tersebut terbagi menjadi dua yaitu *continuous* skip gram dan *continuous bag of words* (CBOW)[19]. Pada penelitian ini digunakan model CBOW di mana model ini mencoba untuk memahami konteks kata dan menjadikannya sebagai masukan dan kemudian mencoba memprediksi kata-kata yang akurat secara kontekstual. Selanjutnya proses ekspansi dilakukan pada semua kata awal yang dianggap mewakili tujuh aspek etik penelitian kesehatan seperti yang tampak pada Tabel 2.

Tabel 2. Daftar kata yang akan diekspansi

No	Aspek Etik	Daftar Kata
1	Sosial	hasil, diseminasi, manfaat, kontribusi
2	Ilmiah	metodologi, prosedur, analisis, operasi
3	Pemerataan beban dan manfaat	distribusi, lokasi, rentan, beban, jaminan
4	Risiko	efek samping, gejala, dilaporkan, kriminal, kehilangan pekerjaan, risiko
5	Bujukan	hadiah, gratis, kompensasi, imbalan, sukarela
6	Privasi	rahasia, identitas, privasi, anonim, kerahasiaan, kode
7	Persetujuan setelah penjelasan	persetujuan, bersedia, paksaan, penjelasan

Tabel 2 merupakan kata atau term awal yang dapat mewakili tujuh aspek etik. Kata-kata tersebut akan melalui proses ekspansi menggunakan *word embedding word2vec*. Hasil proses ekspansi inilah yang nantinya akan dicari nilai kemiripannya dengan hasil proses ekstraksi kata kunci. Sebagai contoh misalnya ada kata "hadiah" pada proses ekspansi pertama kata hadiah menghasilkan kata "bingkisan", "penghargaan", dan "kado".

2.3 Ekstraksi kata kunci

Ekstraksi kata kunci merupakan salah satu bagian dari *text mining* yang fungsinya mengidentifikasi suatu istilah dalam dokumen yang menggambarkan argumen dalam suatu dokumen dengan memanfaatkan algoritma ekstraksi. Selanjutnya dilakukan proses pra pemrosesan proses ekstraksi kata kunci dengan beberapa metode seperti TF-IDF, YAKE dan KeyBERT. TF-IDF merupakan bentuk dari metode representasi teks menjadi numerik. Pada prinsipnya TF-IDF menentukan frekuensi kata dan membandingkannya dengan jumlah atau proporsi kata dalam suatu dokumen. YAKE adalah salah satu metode ekstraksi kata bersifat *unsupervised*. Pendekatan YAKE tidak hanya memanfaatkan proses ekstraksi dari frekuensi kemunculan kata, tetapi juga menghitung awalan kata, konteks kata, posisi serta keberagaman kata pada setiap kalimat. Secara matematis proses YAKE digambarkan dengan persamaan 1.

$$YAKE = \frac{T_{Related}+T_{position}}{T_{case}+\frac{TF_{Normalized}}{T_{Related}}+\frac{T_{Sentence}}{T_{Related}}} \quad (1)$$

Persamaan 1 digunakan untuk melakukan proses perhitungan YAKE di mana $T_{Related}$ adalah Keterkaitan kata dengan konteks, $T_{position}$ merupakan posisi kata, T_{case} jumlah kata

dengan huruf kapital atau akronim, $TF_{Normalized}$ frekuensi kemunculan kata, dan $T_{Sentence}$ yaitu frekuensi kemunculan kata dalam kalimat.

Sementara itu, keyBERT merupakan suatu *library* ekstraksi kata kunci kayyang memanfaatkan *embedding* BERT untuk mendapatkan kata kunci yang paling mewakili dokumen. Sebelum dilakukan proses ekstraksi kata kunci terlebih dahulu dilakukan pembersihan data yang akan ekstraksi. Terdapat beberapa langkah yang digunakan pada membersihkan data pertama adalah *case folding* di mana proses ini merupakan tahapan membuat semua huruf besar (*uppercase*) pada data menjadi huruf kecil (*lowercase*). Selanjutnya yaitu *stopword removal* yaitu proses menghilangkan kata-kata yang dianggap tidak penting atau kata-kata umum seperti kata gabung dan kata tunjuk. Pada proses *stopword removal* digunakan juga kumpulan data *stopword* bahasa Indonesia untuk mempermudah proses menghilangkan kata-kata yang tidak penting. Contoh kata yang akan dibuang seperti kata "yang", "dan", "di", "dari", "adalah", dan lain sebagainya. Setelah itu dilakukan proses *remove punctuation* untuk menghilangkan tanda baca beserta simbol-simbol yang terdapat pada dokumen. Terakhir adalah proses *tokenizing* yang merupakan proses pemisahan setiap kata yang ada pada kalimat. Hasil dari proses ekstraksi kata kunci kemudian akan dilakukan proses pemeringkatan dan 10 kata kunci teratas yang akan dijadikan sebagai kata kunci. Metode pemeringkatan yang digunakan di sini adalah *cosine similarity*. Perhitungan dilakukan dengan mencari nilai kesamaan dengan hasil ekstraksi kata kunci dari masing-masing metode baik TF-IDF, YAKE ,keyBERT atau kombinasi dari ketiga metode tersebut dengan hasil teks aslinya. Sebagai ilustrasi proses ekstraksi kata kunci misalkan sebuah ringkasan penelitian didapatkan hasil sebagai berikut: "penelitian", "infeksi", "studi", "mengevaluasi", "staf", "pengunjung", "penularan", "penyakit", "mencuci", "organisasi". Kata-kata hasil ekstraksi inilah yang akan dicari tingkat kemiripannya dengan hasil kata dan hasil ekspansi dari Tabel 2 di atas.

2.4 Perhitungan Semantic Similarity

Setelah didapatkan hasil penambahan kata baru dari ketujuh aspek etik menggunakan ekspansi kueri dan hasil ekstraksi kata kunci menggunakan tiga pendekatan di atas selanjutnya adalah menghitung *similarity* antara hasil ekstraksi dan hasil penambahan masing-masing ketujuh aspek etik. Untuk menghitung *similarity* pada tahapan ini digunakan *tool* yaitu *wordnet similarity* berbahasa Indonesia. Masing-masing kata dari hasil ekstraksi akan dihitung *similarity* dengan daftar kata baru dari masing-masing aspek baik sebelum dilakukan ekspansi atau setelah dilakukan ekspansi. Kemudian diambil rata-rata dari semua kata yang akan menjadi hasil akhirnya.

$$similarity\ (w_i, w_j) = \frac{\sum_{m-1}^{K} \quad w_i^m . w_j^m}{\sqrt{\sum_{m-1}^{K} \quad (w_i^m)^2}\sqrt{\sum_{m-1}^{K} \quad (w_j^m)^2}} \quad (2)$$

Persamaan 2 digunakan untuk menghitung rata-rata *similarity* dari masing-masing aspek di mana w_i adalah kata 1 dan w_j adalah kata 2 dan $\sum_{m-1}^{K}$ adalah jumlah iterasi m ke kata K. Maksimum nilai similarity distance adalah 1 di mana nilai similarity nya adalah 0 sampai dengan 1. Sebagai contoh kata 1 memiliki *wordnet* (*synset* 1,1 dan *synset* 1.2) sedangkan kata 2 memiliki *wordnet* (*synset* 1,1, *synset* 1.2 dan *synset* 1.3). Ke-dua hasil *wordnet* dibandingkan satu persatu kemudian diambil rata-rata nilai akhirnya.

2.5 Fungsi keanggotaan fuzzy

Setelah mendapatkan hasil nilai *similarity* dari ketujuh aspek etik di atas, langkah selanjutnya adalah melakukan perhitungan dengan menggunakan fuzzy. Ketujuh aspek etik tersebut akan dihitung nilai keanggotaan fuzzy apakah nilai yang dihasilkan masing-masing aspek etik masuk dalam kategori yang telah ditentukan berdasarkan nilai fuzzy *membership function*. Dalam penelitian ini digunakan *triangular*-type untuk mengidentifikasi derajat keanggotaan dari semua variabel *input* dan *output* dari fuzzy. Rentang nilai fungsi keanggotaan memiliki rentan nilai antara 0 sampai 1. Untuk mendapatkan fungsi keanggotaan pada penelitian ini digunakan kurva segitiga. Untuk menghitung fungsi keanggotaan segitiga dapat digunakan Persamaan 3.

$$[x] = \{\ 0 \quad x \leq a\ atau\ x\ (x-a)/(b-a)\ (b-x)/(c-b)\ b \leq x \leq c \quad (3)$$

Di mana pada persamaan tersebut [x] merupakan derajat keanggotaan sedangkan a, b, c adalah batas atas atau bawah dari rentan nilai yang sudah ditentukan. Terdapat tiga fungsi keanggotaan yaitu Tinggi (T) rentan nilai yang digunakan adalah 0,6 – 1. Kemudian untuk *membership* Sedang (S) nilai rentannya adalah antara 0,55 - 0,75. Sedangkan untuk *membership* Rendah (R) nilai rentannya adalah < 0,55. Sebagai contoh terdapat dua kata misalnya "penelitian" dan "metode". Kedua kata tersebut akan dicari synset menggunakan *wordnet* kemudian akan dicari nilai kemiripannya dengan *wup_similarity*. Dari kedua kata tersebut didapatkan hasil Synset ('perusal.n.01') untuk kata "penelitian" dan synset ('method.n.01') untuk kata "metode". Kemudian dua synset tersebut dicari nilai kemiripannya dan didapatkanlah hasil $[x]$ adalah 0.5. Dengan persamaan 1 dan rentan yang sudah dijelaskan di atas maka didapatkan hasil -1 untuk nilai R dan 0.6 untuk nilai S sehingga dapat disimpulkan bahwa derajat keanggotaan $[x]$ 0.5 masuk dalam kategori Sedang (S).

2.6 Metode Evaluasi

Metode evaluasi yang digunakan pada penelitian meliputi pengukuran nilai *precision, recall, dan F1-Score*. Proses perhitungan *precision* dan *recall* ditentukan oleh prediksi informasi pada nilai aktual yang diwakili oleh *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negatif* (FN). *Precision* melakukan perhitungan tingkat akurasi antara informasi yang diminta oleh pengguna dan jawabannya diberikan oleh sistem. Nilai *precision* dapat dihitung dengan menggunakan persamaan 4. *Recall* adalah tingkat keberhasilan sistem di menemukan kembali informasi dan dapat dihitung dengan Persamaan 5. Sedangkan F1-Score adalah nilai keseimbangan antara *precision* dan *recall* dan dapat dihitung dengan Persamaan 6.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1 - Score = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall} \quad (6)$$

Di mana TP merupakan jumlah nilai *true positive* (data prediksi benar sesuai dengan hasil sistem). FP adalah nilai *false positive* (data salah yang diperdiksi benar oleh sistem) dan FN adalah nilai *false negative* (data benar yang diprediksi salah oleh sistem).

3. HASIL DAN PEMBAHASAN

Studi ini merupakan studi tentang deteksi etik pada penelitian kesehatan. Pada penjelasan sebelumnya pada penelitian ini menggunakan empat metode evaluasi untuk mengukur kinerja metode yang diusulkan seperti *Precision, Recall,* Dan *F1-Score*. Beberapa percobaan yang dilakukan untuk menguji metode yang diusulkan. Uji coba pertama adalah mencoba membandingkan hasil dari metode ekstraksi kata kunci ekstraksi kata kunci tanpa melakukan ekspansi kueri. Artinya penentuan aspek etik dari ekstraksi kata kunci hanya akan dicari tingkat kemiripannya dengan kata awal saja yang terdapat pada Tabel 2. Uji Coba kedua merupakan proses melakukan deteksi etik dari hasil kata kunci dengan melakukan proses ekspansi kueri di mana ekspansi kueri (QE1) dilakukan untuk memperluas kata yang ada pada Tabel 2. Selanjutnya uji coba ketiga merupakan proses melakukan deteksi etik dari hasil kata kunci dengan melakukan proses ekspansi kueri pada hasil ekspansi sebelumnya artinya hasil ekspansi yang pertama dilakukan proses ekspansi kembali (QE2). Semua proses uji coba selanjutnya akan dibandingkan performanya dengan pendekatan fuzzy.

Tabel 3 Hasil Uji Coba Tanpa

Menggunakan Fuzzy

Metode	TANPA QE			DENGAN QE 1			DENGAN QE 2		
	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
TFIDF	0.42	0.56	0.48	0.45	0.57	0.51	0.43	0.61	0.51
YAKE	0.41	**0.64**	**0.50**	0.45	**0.69**	**0.54**	**0.43**	**0.72**	**0.54**
keyBERT	0.40	0.53	0.46	0.44	0.56	0.50	0.40	0.58	0.47
TFIDF + YAKE	**0.43**	0.65	0.52	0.44	0.66	0.53	0.44	0.69	0.54
TFIDF + keyBERT	**0.43**	0.56	0.49	0.46	0.57	0.51	0.43	0.60	0.50
YAKE + keyBERT	0.42	0.56	0.49	0.45	0.57	0.50	0.41	0.61	0.49
TFIDF+YAKE+ keyBERT	0.42	0.57	0.48	**0.46**	0.58	0.51	**0.43**	0.61	0.50

Tabel 3 menunjukkan hasil dari percobaan pertama dan kedua yaitu mencoba masing-masing metode ekstraksi kata kunci seperti TF-IDF, YAKE, BERT dan kombinasi dari tiga metode tersebut baik yang dengan atau tanpa proses ekspansi pada kata awal. Hasilnya menunjukkan keberagaman nilai. Secara *precision* model YAKE dan gabungan antara TFIDF+YAKE+keyBERT relatif unggul jika dibandingkan dengan metode lainnya dan untuk *recall* dan f1-score model YAKE juga terbilang unggul jika dibandingkan dengan metode lain. Jika dilihat dari sebaran hasil maka metode YAKE dapat dikatakan unggul dari sisi *recall* dan rata-rata *recall* dan *precision*. Selain itu pada Tabel 3 juga terlihat perbandingan hasil antara yang menggunakan atau tanpa proses ekspansi menunjukkan tren peningkatan hasil. Namun pada ekspansi yang kedua terjadi penurunan hasil jika dibandingkan dengan ekspansi pertama walaupun masih tetap unggul jika dibandingkan dengan yang tidak dilakukan ekspansi. Hal ini menjuakkan semakin banyak dan relevan kata yang dijadikan sebagai acuan maka semakin bagus angka kemiripannya walaupun dalam proses perhitungan kemiripan dengan menggunakan *wordnet* bahasa Indonesia banyak term atau istilah yang tidak terdapat dalam kumpulan data *synset*.

Tabel 4 Hasil Uji Coba Menggunakan Fuzzy

Metode	TANPA QE			DENGAN QE 1			DENGAN QE 2		
	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
TFIDF	0.42	0.54	0.47	0.48	0.54	0.51	0.43	0.54	0.48
YAKE	0.42	**0.58**	**0.50**	0.47	**0.61**	**0.53**	0.44	**0.66**	**0.52**
keyBERT	0.41	0.52	0.46	0.44	0.50	0.47	0.39	0.51	0.44
TFIDF + YAKE	0.43	0.57	0.51	0.45	0.58	0.51	0.45	0.58	0.51
TFIDF + keyBERT	0.43	0.53	0.47	0.47	0.49	0.48	0.43	0.52	0.47
YAKE + keyBERT	0.42	0.56	0.48	0.47	0.53	0.50	0.43	0.56	0.48
TFIDF+YAKE+ keyBERT	**0.44**	0.53	0.47	**0.49**	0.53	0.51	**0.59**	0.43	0.50

Tabel 4 menunjukkan hasil percobaan yang menggunakan fuzzy. Jika dilihat dari nilai *precision* hasil tidak jauh berbeda dengan yang hasil percobaan sebelumnya tanpa menggunakan fuzzy. Secara *precision* kombinasi TFIDF+YAKE+ keyBERT masih unggul jika dibandingkan dengan yang lainnya. Sementara itu, dari evaluasi *recall* metode YAKE memiliki nilai yang relatif lebih tinggi dibandingkan dengan metode lainnya. Jika dibandingkan antara proses yang menggunakan fuzzy dan tidak, maka penggunaan fuzzy hanya mampu meningkatkan nilai *preccision* namun dari segi *recal* dan *f1-score* hasilnya cenderung mengalami penurunan.

Dilihat dari sisi ekstraksi kata kunci baik dalam percobaan yang menggunakan fuzzy ataupun tanpa fuzzy kombinasi beberapa metode ekstraksi tidak memberikan dampak yang signifikan. Kombinasi metode ekstraksi kurang memberikan dampak terhadap peningkatan hasil dari tiga metode evaluasi yang digunakan. Jika ditelusuri lebih jauh, kata kunci yang dihasilkan dari gabungan metode ekstraksi tidak jauh lebih baik dari hasil metode ekstraksi yang dihasilkan tanpa kombinasi karena beberapa metode seperti keyBERT hasil kata kunci yang didapatkan tidak relevan dengan aspek etik. Sebagai contoh kata kunci yang dihasilkan keyBERT dalam beberapa dokumen adalah seperti kata-kata berikut "ketidakseragaman", "ketidaklengkapan", "mengoptimalkan". Kata kunci tersebut tentu tidak memiliki relevansi dengan aspek etik. Selain tidak memiliki relevansi kata kunci tersebut juga tidak ditemukan synset-nya pada *wordnet* sehingga kata kunci tersebut tingkat kesamaannya dengan aspek etik adalah 0. Jika dibandingkan dengan kata dengan hasil ekspansi dengan metode lainnya hasil kata kunci yang dihasilkan keyBERT banyak ditemukan tidak memiliki relevansi dengan aspek etik.

Penggunaan ekspansi menjadi temuan dan kelebihan tersendiri dalam penelitian ini karena mampu memberikan tambahan nilai dari hasil evaluasi yang dilakukan dengan tiga metode evaluasi di atas. Artinya, semakin banyak dan semakin relevan hasil ekspansi maka semakin tinggi tingkat kemiripan dengan kata kunci yang dihasilkan dari ringkasan. Penggunaan *word2vec* memiliki kelebihan yaitu dapat mengembalikan kata yang memiliki nilai *co-occurrence* yang tinggi. Sebagai contoh yang kata yang berhasil adalah kata "hadiah" jika dilakukan proses ekspansi maka hasil ekspansinya adalah "hadiahnya, bingkisan, penghargaannya, penghargaan, trofi, mas kawin, beasiswa, kado dan sumbangan". Berdasarkan kata hasil ekspansi kata yang dihasilkan dari hadiah dinilai cukup relevan dengan aspek etik meskipun terdapat beberapa kata kalau dikaitkan dengan konteks etik bujukan tidak memiliki relevansi seperti kata "trofi". Hal ini terjadi karena dalam proses *word2vec* menggunakan model umum yaitu library *word2vec* umum wikipedia bahasa Indonesia sehingga potensi biasanya cukup tinggi. Selanjutnya berdasarkan hasil percobaan yang dilakukan hasil QE1 lebih baik dari QE2 ini disebabkan karena beberapa hasil ekspansi tahap kedua tidak ditemukan synset-nya dalam *wordnet*. Selain itu juga terdapat beberapa dari hasil ekspansi kedua tingkat kemiripan katanya dengan hasil kata kunci kecil sehingga hal tersebut tidak masuk dalam perhitungan karena dalam proses fuzzy yang diambil hanya yang bernilai sedang dan tinggi dengan rentan 0.58 sampai dengan 1.

4. KESIMPULAN DAN SARAN

Dalam makalah ini diusulkan gabungan beberapa metode yaitu ekstraksi kata kunci, ekspansi kueri dan perhitungan fuzzy untuk mendeteksi adanya aspek etik pada ringkasan penelitian kesehatan. Pendekatan TF-IDF, YAKE, KeyBERT dan kombinasi ke tiganya digunakan untuk melakukan ekstraksi kata kunci. Sementara untuk ekspansi kueri digunakan pendekatan *word embeding word2v*ec. Hasil percobaan yang telah dilakukan dan diukur dengan tiga model evaluasi menunjukkan bahwa pendekatan TF-IDF dengan dua kali ekspansi kueri mendapat nilai *precision* paling tinggi yaitu 70%. Sementara secara keseluruhan metode YAKE mendapat nilai tertinggi dengan angka 51%. Penggunaan fuzzy berdasarkan hasil uji coba yang telah dilakukan hanya mampu meningkatkan nilai *recall*. Penggunaan *wordnet* berbahasa Indonesia menjadi keterbatasan tersendiri dalam penelitian ini karena beberapa dari kata hasil ekstraksi dan kata kunci tidak dapat ditemukan hasil synsetnya sehingga ke depan perlu ada pendekatan lain seperti penerapan metode yang mampu mengekstraksi konteks kalimat positif dan negatif penggunaan model lain selain penggunaan ekstraksi kata kunci seperti klasifikasi.

DAFTAR PUSTAKA

[1] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang,J., 2020, BioBERT: A pre-trained biomedical language representation model for biomedical text mining, Bioinformatics, no. 4, vol. 36, 1234–1240, 2020.

[2] Kementerian Kesehatan RI, 2017, Pedoman dan standar etik penelitian dan pengembangan kesehatan nasional," Kementerian Kesehatan RI, 1–158.
[3] Handayani, L. T., 2018, Kajian etik penelitian dalam bidang kesehatan dengan melibatkan manusia sebagai subyek, The Indonesian Journal Of Health Science, no. 1, vol. 10,47–54, 2018.
[4] Budi S., 2017, Text mining untuk analisis sentimen review film menggunakan algoritma k-means," Techno.Com, no. 1, vol. 16, 1–8.
[5] Wahyudi, M. D. R., 2019, Penerapan Algoritma Cosine Similarity pada Text Mining Terjemah Al-Qur'an Berdasarkan Keterkaitan Topik, Semesta Tek., no. 1, vol. 22, 41–50.
[6] Pavlopoulos, I.J., 2014, Aspect based sentiment analysis Ioannis, Thesis, Department Of Informatics, Athens University Of Economics And Business, Athena.
[7] Priyantina, R.A., Sarno, R., 2019, Sentiment analysis of hotel reviews using latent dirichlet allocation, semantic similarity and LSTM, International Journal of Intelligent Engineering and System, no. 4, vol. 12, 142–155.
[8] Rose, S., Engel , D., Cramer , N., Cowley, W., 2010, Automatic keyword extraction, Text Min. Appl. Theory, pp. 1-277.
[9] Chamidah, N., Santoni, M., 2021, Pencocokan berbasis kata kunci pada penilaian esai pendek otomatis berbahasa indonesia, Techno.Com, no. 1, vol. 20, 19–27.
[10] Koloski, B., Pollak, S., Skrlj, B., Martinc, M., 2021, Extending neural keyword extraction with tf-idf tagset matching, *EACL Hackashop News Media Content Anal. Autom. Rep. Gener. Hackashop 16th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL*.
[11] Campos, R., Mangaravite, V., Pasquali A., Jorge,A., Nunes, C.,Jatowt ,A., 2020, YAKE: Keyword extraction from single documents using multiple local features," Inf. Sci. (Ny)., vol. 509, 257–289.
[12] Campos, R., Mangaravite, V., Pasquali A., Jorge,A., Nunes, C.,Jatowt ,A., 2018, A text feature based automatic keyword extraction method for single documents," Lect. Notes Comput. Sci, no. 1, vol. 10772 LNCS,684–691, 2018.
[13] Priyanshu, A., Vijay, S., 2022, AdaptKeyBERT: An Attention-Based approach towards Few-Shot & Zero-Shot Domain Adaptation of KeyBERT, *arXiv preprint:* https://arxiv.org/pdf/2211.07499
[14] Perez, G.Y., A. Simon, C.W., Hojas, M.J., Olivas, A., Romero, F. P., 2018, A fuzzy approach to improve an unsupervised automatic keyphrase extraction process, *IEEE Int. Conf. Fuzzy Syst*, July.
[15] Sheeba, J. I., 2014, A fuzzy logic based improved keyword extraction from meeting transcripts, *International Journal on Computer Science and Engineering*, no. 08, vol. 6, 287–299.
[16] M. A. Raza, R. Mokhtar, and N. Ahmad, 2019, A survey of statistical approaches for query expansion, *Knowl. Inf. Syst.*, vol. 61, no. 1, hal. 1–25.
[17] Maryamah, Arifin, A.Z., Sarno, R., Indraswari, R.R., Sholikah, W.,2021, Pseudo-relevance feedback combining statistical and semantic term extraction for searching arabic documents, *International Journal of Intelligent Engineering and System*, no. 5, vol. 14, 238–246.
[18] Diaz, F., Mitra, B., Craswell, N., 2016, Query expansion with locally-trained word embeddings, *54th Annu. Meet. Assoc. Comput. Linguist.*
[19] Kuzi, S., Shtok, A., Kurland, O., 2016, Query expansion using word embeddings, *Int. Conf. Inf. Knowl. Manag. Proc.*, October-24.